МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО”

Сергеєв Данило Сергійович

УДК 004.93

Інформаційна технологія обробки природномовних текстів на основі інтеграційного підходу

Спеціальність: 05.13.06 – інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Київ - 2019

Дисертацією є рукопис.
Робота виконана на кафедрі технічної кібернетики Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського” Міністерства освіти і науки України.

Науковий керівник: кандидат технічних наук, доцент
Кисленко Юрій Іванович,
Національний технічний Університет України
“Київський політехнічний інститут імені Ігоря Сікорського”,
доцент кафедри технічної кібернетики

Офіційні опоненти: доктор технічних наук, професор
Бармак Олександр Володимирович,
Хмельницький національний університет,
професор кафедри комп’ютерних наук та інформаційних технологій

доктор технічних наук, професор
Ланде Дмитро Володимирович,
Інститут проблем реєстрації інформації Національної академії наук України,
завідувач відділу спеціалізованих засобів моделювання

Захист відбудеться «04» жовтня 2019 р. о 15 годині 00 хвилин на засіданні спеціалізованої вченої ради Д 26.002.29 при Національному технічному університеті України «Київський політехнічний інститут імені Ігоря Сікорського» за адресою: 03056, м. Київ, просп. Перемоги, 37, корп. №11, ауд.215.

З дисертацією можна ознайомитися у Науково-технічній бібліотеці імені Г.І. Денисенка Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» за адресою: 03056, м. Київ, просп. Перемоги, 37.

Автореферат розісланий «04» вересня 2019 р.

Учений секретар
спеціалізованої вченої ради _____ С.Ф. Теленик

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Обробка природної мови (ОПМ) – загальний напрямок інформатики, штучного інтелекту та математичної лінгвістики, який вивчає проблеми комп'ютерного аналізу та синтезу природної мови. Прикладні задачі ОПМ охоплюють багато підгалузей, зокрема виділення інформації, природномовний пошук, машинний переклад, ідентифікація плагіату та багато інших.

За останні роки дослідження у галузі ОПМ досягли значних практичних результатів, зокрема досягнений суттєвий прогрес у розробці природномовного голосового інтерфейсу користувача для мобільних пристроїв, технологіях машинного перекладу, технологіях розпізнавання рукописного тексту та голосу тощо. При цьому актуальною залишається задача покращення якості роботи прикладних систем, які реалізують ці технології. Так, дослідження показують, що складніші прикладні технології ОПМ, зокрема системи машинного перекладу та природномовного пошуку, показують гірші результати ніж окремі технології нижчих рівнів, і є потенціал для їх удосконалення на основі існуючого стеку технологій. Ключовим елементом такого удосконалення постає використання баз знань, які одночасно взаємодіють з прикладними технологіями ОПМ різних рівнів, і зокрема природномовних баз знань.

Основи досліджень в цій галузі заклали такі вчені як E.D. Liddy, J.F. Sowa, R. Harris, K. Brown, R. Tadeusiewicz, N. Chomsky, D. Herrmann, N. Chapman, J. Lyons, О.В. Бармак, М.З. Згуровський, О.А. Кришталь, Д.В. Ланде, В.А. Лефевр, В.М. Томашевський, В.А. Широков.

Природномовні бази знань (ПМБЗ) – це клас баз знань, об’єктом роботи яких є природномовна інформація – тобто, знання в яких зберігається безпосередньо у вигляді природномовної інформації, на відміну від інших класів баз знань, у яких природна мова зберігається з використанням спеціалізованих формальних моделей представлення знань. Основним напрямком сучасних досліджень в галузі ПМБЗ є розробка гібридних ПМБЗ, що поєднують різні моделі ПМБЗ, як-то мережева модель знань з частково рекурсивними елементами. Втім, такі системи часто наслідують не лише переваги, але й недоліки систем, на яких вони засновані. Крім того, розробка гібридної ПМБЗ є технічно складною задачею, і така система часто обмежена умовами прикладної задачі, для вирішення якої вона створюються – а саме, може містити принципові недоліки, які не впливають на вирішення окремої задачі, але є суттєвими для ПМБЗ загального використання.

Значний внесок в розвиток цих ідей зробили M. Weigt, C. Baker, B. Cronin, R. Brachman, G. Antoniou, C. Fillmore, J.M. Hellerstein, M. Blanton, A. Pable, V.E. Wolfengagen, А. Левицький, В.В. Бочкаров, А.В. Анісімов, П.І. Федорчук, О.В. Іванов та інші. Розроблений на основі робіт цих вчених інтеграційний підхід дозволяє вирішити деякі з відомих проблем комп’ютерної лінгвістики, а розроблені на його засадах природномовні бази знань надають нові можливості для їх використання в технологіях обробки природної мови.

Перевагами існуючих підходів є їх висока ефективність у вирішенні відповідних спеціалізованих задач, як-то розпізнавання символів або статистичного аналізу текстів. Водночас, суттєвими їх недоліком є відсутність системної взаємодії

між технологіями різних рівнів, що призводить до недостатньо ефективної роботи комплексних технологій повного циклу обробки природної мови, зокрема систем природномовного пошуку та машинного перекладу.

Наукове завдання дослідження полягає у розробці інформаційної технології обробки природномовних текстів на основі інтеграційного підходу з метою підвищення ефективності роботи технологій обробки природної мови.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у відповідності з планами науково-дослідницької роботи кафедри технічної кібернетики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» в рамках НДР «Проектування лінгвістичного процесора та бази знань як складових індивідуальної мовної системи для формування цілого кластеру інформаційних природномовних технологій» (номер держреєстрації 0117U004327), в якій автору належить формалізована модель природномовних знань та структура моделі світу на її основі.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності обробки природномовної інформації за рахунок використання гнучкої моделі представлення природномовних знань шляхом розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

Об'єкт дослідження – процес обробки природномовної інформації.

Предмет дослідження – моделі, методи, алгоритми та інформаційні технології обробки природномовної інформації.

Відповідно до мети були поставлені і виконані наступні завдання.

1. Проаналізовано існуючі технології обробки природної мови та виконано аналіз особливостей використання природномовних баз знань у цих технологіях.
2. Розроблено модель представлення природномовних знань на основі інтеграційного підходу, що враховує особливості природної мови як об'єкту роботи бази знань.
3. Розроблено метод виділення знань з природномовного тексту та пошуку знань з урахуванням розробленої моделі представлення знань.
4. Розроблено інформаційну технологію обробки природномовних текстів на основі моделі представлення знань та методу їх обробки.
5. Виконано експериментальну перевірку розробленої інформаційної технології.

Методи дослідження. У даній роботі були використані такі методи: методи розробки баз даних та баз знань; метод ієрархічної декомпозиції; методи системного аналізу; аксіоматичний метод у частині вихідних теоретичних положень; метод аналітичної оцінки обчислювальної складності алгоритму; методи моделювання та експерименту для перевірки створеної інформаційної технології.

Наукова новизна одержаних результатів.

1. Вперше запропоновано модель представлення природномовних знань на основі інтеграційного підходу до моделювання мовленнєвої діяльності людини, в основі якої лежить формалізоване представлення кванту знань у вигляді фрагменту довільного природномовного тексту, що дозволяє зберігати структуру тексту у вигляді однотипних структур даних.

2. Вперше розроблено метод опрацювання природномовних текстів з використанням запропонованої моделі представлення природномовних знань, який виділяє з тексту складові окремих квантів знань, що дозволяє виділити структуру знань довільного природномовного тексту.
3. Вперше розроблено інформаційну технологію обробки природномовних текстів на основі використання запропонованої моделі представлення знань та методу опрацювання природномовних текстів, що дозволяє виконувати повнотекстовий природномовний пошук за логарифмічний обчислювальний час.

Практичне значення одержаних результатів.

1. Розроблений в рамках даної роботи метод обробки природномовних текстів може бути використаний для покращення роботи технологій обробки природної мови, зокрема систем природномовного пошуку в мережі Інтернет, систем машинного перекладу, природномовних інтерфейсів користувача.
2. Створена в рамках даної роботи модель представлення природномовних знань може бути використана для розробки універсальних природномовних баз знань.
3. Окремі результати розробленої в рамках даного дослідження інформаційної технології було впроваджено в робочий процес ТОВ «Діджитал принт» та ТОВ «Міжнародна текстильна корпорація».
4. Розроблені в рамках даного дослідження моделі і методи впроваджено в матеріалах курсів «Теорія алгоритмів» та «Візуальне програмування» у навчальний процес кафедри технічної кібернетики факультету інформатики та обчислювальної техніки КПІ ім. Ігоря Сікорського.

Особистий внесок здобувача. Усі результати, що складають основний зміст дисертаційної роботи, отримані автором самостійно.

У спільних роботах автору належить: модель бази знань, функціональна модель інтерфейсу у *[Cognitive architecture of speech activity and modelling thereof, BICA, 2015]*; методика аналізу та опрацювання даних у *[Структурний підхід до пошуку природно-мовної інформації, Радіоелектроніка та інформатика. 2015]*; формалізація проблеми, формалізація використаного підходу, збір та опрацювання даних у *[Порівняння способів збереження слів в IT", АСАУ, 2016];* формалізація та обґрунтування використаної моделі у *[Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях", АСАУ, 2016]*.

Апробація результатів дисертації. Основні результати роботи доповідалися та обговорювалися на:

міжнародній конференції «Електроніка та інформаційні технології / ЕлІТ-2015» з темою доповіді «Особливості моделювання бази природно-мовних знань» (Львів-Чинадієво, 27-30 серп. 2015 р.);

міжнародній конференції «System Analysis and Information Technology / SAIT-2016» з темою доповіді «Методика оцінки якості природно-мовних пошукових систем»;

міжнародній конференції «Електроніка та інформаційні технології / ЕлІТ-2016» з темою доповіді «Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції» (Львів-Чинадієво, 27-30 серп. 2016 р.);

міжнародній конференції «Штучний інтелект та інтелектуальні системи AIIS'2016» з темою доповіді «Комп'ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань» (Київ, 29.11-2.12.2016 р.);

міжнародній конференції «System Analysis and Information Technology / SAIT-2017» з темою доповіді «Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань» (Київ, 22.05-25.05.2017 р.).

Публікації. За результатами досліджень опубліковано 12 наукових праць, у тому числі 6 статей у наукових фахових виданнях (з них 1 стаття у виданнях іноземних держав, 4 у наукових фахових виданнях України, які входять до міжнародних наукометричних баз), 6 тез доповідей в збірниках матеріалів конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, списку використаних джерел (132 найменування) та 6 додатків. Загальний обсяг роботи складає 165 сторінок. Основна частина дисертації займає 118 сторінок, містить 28 рисунків та 12 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтована актуальність теми дисертаційної роботи, сформована мета, ідея і задачі дослідження, наукова новизна й практичне значення отриманих результатів, наведені наукові положення, що виносяться на захист, наукове та практичне значення роботи, дані про публікації, апробацію та впровадження розробок і результатів дослідження.

У **першому розділі** виконано аналіз задач та процесів комп'ютерної обробки природної мови, проаналізовано прикладні аспекти використання технологій обробки природної мови в інформаційних технологіях. Визначено роль баз знань в інформаційних технологіях обробки природної мови як компонента, необхідного для взаємодії різних їх систем, охарактеризовано та проаналізовано існуючи підходи до проектування природномовних баз знань.

Показано, що актуальною є задача розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

У другому розділі визначено особливості розробки природномовної бази знань для інформаційної технології обробки природномовних текстів на основі інтеграційного підходу. З урахуванням цих особливостей створено формальну модель представлення знань у природномовній базі знань та розроблено моделі її основних елементів, якими є квант знань, або найменший елемент знань, та відношення, яке описує зв'язок між квантами знань. З використанням моделі представлення знань створено метод обробки природномовних текстів для інформаційної технології та процедури використання інформаційної технології у прикладних задачах обробки природної мови.

З урахуванням принципів інтеграційного підходу (ІП) побудовано структурну схему природномовної бази знань для використанні у складі інформаційної технології обробки природномовних текстів на основі інтеграційного підходу (ПМБЗ ІП). Визначено, що підсистеми ПМБЗ у випадку ПМБЗ ІП виходять за рамки ПМБЗ і на рівні самої ПМБЗ представлені лише інтерфейсами обміну запитами.

Найменшим елементом ПМБЗ ІП є квант знань (КЗ) – двоскладова структура, яка включає в себе суб'єкт $Subj$, предикатор Mov та їх атрибутивне оточення, а саме атрибути $Attr(Subj)$ і $Attr(Mov)$ та міри цих атрибутів $Attr(Attr(Subj))$ і $Attr(Attr(Mov))$.

Загальна структура КЗ має наступний вигляд:

$$Q = \begin{Bmatrix} S, Attr(S) \\ P, Attr(P) \end{Bmatrix}, \quad (1)$$

де Q – квант знань.

S та P – суб'єкт та предикатор даного КЗ;

$Attr(S)$ та $Attr(P)$ – атрибутивні оточення суб'єкта та предикату.

Структура КЗ поєднана з його текстовим представленням через лексеми.

Лексема – це окреме слово з усією сукупністю властивих йому форм словозміни й значень у різних контекстах. В контексті ПМБЗ ІП це означає, що усі граматичні та смислові форми окремого слова не є самостійними сутностями, а належать до єдиного об'єкту.

Лексема є структурою наступного вигляду:

$$L = \{l_0, [l_1, \dots, l_i, \dots]\}, \quad (2)$$

де l_0 – базова форма слова;

l_i – (за наявності) – інші граматичні форми слова.

Використання такої структури даних дозволяє прив'язати одну й ту ж саму структуру знань до різних варіантів її представлення на рівні тексту, уніфікувати структурне представлення елементів КЗ ПМБЗ та також правильно обробляти ситуації, коли певна форма лексеми на рівні тексту представлена кількома словами.

Кожний з елементів КЗ може складатися з декількох лексем.

Структура $Subj$ з урахуванням лексем виглядає наступним чином:

$$Subj = \{L_i\},\ i = \overline{1, N_l^S}, \quad (3)$$

де L_i – лексеми, що формують текстове представлення $Subj/Obj$;

N_l^S – кількість таких лексем.

Структура $Pred$ з урахуванням лексем виглядає наступним чином:

$$Pred = \{L_j\},\ j = \overline{1, N_l^P}, \quad (4)$$

де L_j – лексеми, що формують текстове представлення $Pred$;

N_l^P – кількість таких лексем.

Об'єкт $Attr(X)$, що описує атрибутивне оточення деякого елементу X у КЗ, складається з множини атрибутів та мір атрибутів елементу X в рамках даного КЗ.

Цей об'єкт виглядає наступним чином:

$$Attr(X) = A_j,\ j = \overline{1, N_{attr}^X}, \quad (5)$$

де A_j – об'єкт, що описує окремий атрибут A;

N_{attr}^X – загальна кількість атрибутів X, певного елементу КЗ.

Загальна структура атрибута A_j виглядає наступним чином:

$$A_j = \{a, M(a)\}, \quad (6)$$

де a – власне атрибут;

$M(a)$ – множина мір атрибуту a.

Множина мір атрибута a виглядає наступним чином:

$$M(a) = m_k,\ k = \overline{1, N_m^a}, \quad (7)$$

де m – окрема міра атрибуту a;

N_m^a – загальна кількість мір атрибуту a для даного елементу X.

Кожному з атрибутів a та їх мірам m_{Ai} також можемо поставити у відповідність лексеми L, які представляють їх на текстовому рівні, аналогічно $Subj$ та $Pred$.

Окремі КЗ поєднані у спільну структуру знань за допомогою відношень. Відношення – це логічні зв'язки між окремими КЗ у ПМБЗ.

Формула відношення має наступний вигляд:

$$R = \{Q_1, Q_2, \{g, t, d\}\}, \quad (8)$$

де Q_1, Q_2 – це КЗ, між якими встановлено відношення;

$g \in D$ – це граматична конструкція або правило, яке ідентифікує дане відношення R у тексті;

t – це інформація про тип відношення;

$d \in D$ – інформація про напрям відношення.

Фрагмент знань ПМБЗ складається з сукупності КЗ та відношень між ними. Для довільного фрагменту тексту T структура знань приймає наступний вигляд:

$$K_T = \{(Q_1, Q_2, \dots, Q_i, \dots), (R_1, R_2, \dots, R_j, \dots)\}, \quad (9)$$

де T – фрагмент тексту;

K – фрагмент знань, що відповідає фрагменту тексту T;

Q_i – КЗ, що належать до K;

R_j – відношення між КЗ у K.

Довільний фрагмент знань ПМБЗ ІП представляє собою об'єктний граф, що має таку структуру:

$$K = (Q, R), \quad (10)$$

де Q – множина КЗ у ПМБЗ ІП, які є вершинами графу знань;

R – множина відношень ПМБЗ ІП, які є ребрами графу знань.

Фрагмент знань K_T, який описує структуру знань фрагменту тексту T, що є частиною наповнення ПМБЗ ІП – це зв'язний граф, що є підграфом загального графу знань ПМБЗ:

$$K_T \subseteq K. \quad (11)$$

У роботі також розроблено процедуру виділення КЗ та відношень з тексту.

Процедура виділення КЗ з тексту має наступний вигляд:

1. Знайти усі лексеми, що відповідають елементам КЗ.
   1.1. Знайти лексеми, що мають граматичний клас «іменник» ($Subj$).
   1.2. Знайти лексеми, що мають граматичний клас «дієслово» ($Pred$).
   1.3. Знайти лексеми, що мають граматичний клас «прикметник» та «прислівник» ($Attr, Attr(Attr)$).
2. Визначити можливі варіанти формування КЗ.
   2.1. Для кожного $Subj$:
      2.1.1. Для кожного $Pred$, що може бути пов'язаний з цим $Subj$, створити новий КЗ.
      2.1.2. Якщо немає жодного $Pred$, створити новий КЗ з даним $Subj$ та $Pred$ «є».

2.2.Для кожного $Attr$:

2.2.1. Додати до цього $Attr$ усі $Attr(Attr)$, що можуть бути з ним пов'язані.

2.2.2. Додати цей $Attr$ до кожного КЗ, до якого він може належати.

3. Повернути усі сформовані КЗ.

Процедура виділення відношень з тексту має наступний вигляд:

1. Знайти усі граматичні конструкції, що відповідають відомим відношенням
2. Для кожного відношення знайти відповідні йому КЗ:
   2.1.Якщо не знайдено жодного КЗ, ігнорувати відношення
   2.2.Якщо знайдено один КЗ, побудувати відношення з посиланням на поточну ситуацію
   2.3.Якщо знайдено два КЗ, побудувати повне відношення
   2.4.Якщо знайдено більше, ніж два КЗ, побудувати повне відношення з кожним з них
3. Додати кожне відношення у мережу КЗ як об'єкт, що поєднує відповідні КЗ
4. Повернути множину КЗ та відношень

На основі моделі та процедур розроблено метод обробки природномовних текстів. Цей метод включає у себе два класи задач – задачі синтезу («знання-текст») та аналізу («текст-знання»). Для їх вирішення використовується ПМБЗ ІП, яка оперує незалежними об'єктами – природномовним текстом та знаннями. Основною з цих двох задач є задача аналізу, оскільки при цьому відбувається виділення знання з тексту, який може не містити усієї необхідної для цього інформації. Задача обробки природної мови при перетворенні «текст-знання» включає наступні етапи.

1. Синтаксичний аналіз: перетворення тексту в формалізований вигляд. Задача, яка вирішується на цьому етапі, це нормалізація текст T_0 до тексту T, тобто виділення змістовних елементів та формування на основі них синтаксичної структури S, що включає в себе слова як елементи тексту, їх граматичні форми, порядок у тексті та синтаксичні зв'язки та граматичні ролі.

$$T_0 \to T, T \to S, \quad S: (L, L_i, d). \tag{12}$$

Отримана в результаті синтаксична структура S містить так дані:

– лексеми L, в тому числі їх граматичні форми;
– індекси слів w_i, які визначають порядок даного слова у вхідному тексті;
– синтаксичні зв'язки між словами, які визначають належність слів до однієї з груп ($Noun, Verb, Adj, Adv$).

З метою прив'язки фрагменту вхідного тексту T_0 до отриманої з нього синтаксичної структури S, до кожної структури S додається маркер M, який пов'язує слова вхідного тексту w та лексеми синтаксичної структури L.

$$M(T, S): \quad T(w, w_i) \leftrightarrow M \leftrightarrow S(L, L_i). \tag{13}$$

Це дозволяє формально поєднати елементи вхідного тексту та відповідні їм лексеми, які є елементами мережі знань.

2. Виділення КЗ та відношень: створення та наповнення мережі знань на основі отриманої раніше структури S. Мета цього етапу – сформувати кванти знань та

відношення, які відтворюють структуру знань даного природномовного тексту, та заповнити їх з отриманої раніше синтаксичної структури.

На основі лексем з синтаксичної структури S лінгвістичний процесор формує кванти знань, в кожному з яких $Noun$ має роль $Subj$, а $Verb$ – роль $Pred$.

$$Q = \{Subj(Noun)\},$$
$$if\,(form(Verb) = form(Noun)) : Q \leftarrow Verb, \quad (14)$$
$$Q = \{Subj(Noun), Pred(Verb)\}.$$

Доповнення КЗ атрибутивним оточенням відбувається аналогічно.

Таким чином, на основі лексем L синтаксичної структури S, які мають граматичні форми $Noun, Verb, Adj, Adv$, формується повний квант знань:

$$Q = \begin{Bmatrix} Subj(Noun), \\ Attr(Subj)(Adj), \\ Attr\big(Attr(Subj)\big)(Adv), \\ Pred(Verb), \\ Attr(Pred)(Adv) \\ Attr(Attr(Pred))(Adv) \end{Bmatrix}. \quad (15)$$

При виникненні багатозначності, так само як і у випадку $Subj - Pred$ усі варіанти формування КЗ з тексту зберігаються як конкурентні варіанти.

Аналогічним чином відбувається формування відношень:

$$R = \{Q_1, Q_2, \{g, t, d\}\}. \quad (16)$$

3. Формування мережі знань та зв'язків між текстом та знаннями.

З існуючих елементів формується єдина мережа знань, яка включає в себе по одному разу кожну унікальну лексему, кожний унікальний КЗ та кожне унікальне відношення. Результатом виконання цього кроку є мережа знань, структура якої відповідає моделі представлення знань ПМБЗ. Отримана мережа знань прив'язується на структурному рівні до визначених раніше структур та інших джерел інформації – синтаксичної структури вхідного тексту, окремих квантів знань, документів, додаткової семантичної інформації тощо за допомогою ієрархічної структури маркерів $M_Q - M_R - M_T$ (квантів знань, відношень та фрагментів тексту відповідно).

Отримана мережа знань має високу повноту, гнучкість та несуперечливість.

У третьому розділі розроблено структурну схему інформаційної технології обробки природномовних текстів на основі інтеграційного підходу (далі – ІТ ОПМ), виконано аналіз процесів обробки даних в інформаційній технології, зокрема визначено етапи роботи інформаційної технології в режимах аналізу та синтезу та розроблено процедури записування та пошуку природномовних знань у базі знань у складі інформаційної технології. Також наведено приклади використання інформаційної технології обробки природномовних текстів у прикладних задачах природномовного пошуку та машинного перекладу.

Розроблена інформаційна технологія має такі особливості, які визначають її місце серед інших технологій цього класу:

– вхідними даними є природномовний текст;
– обробка тексту виконується згідно методу, описаному у підрозділі 2.5;

– результатом роботи ІТ ОПМ є природномовні знання, модель яких відповідає моделі, описаній у підрозділі 2.3;

– результати роботи ІТ ОПМ орієнтовані на подальшу обробку іншими системами ОПМ.

Структурно ІТ ОПМ складається з окремих підсистем, представлених на рис. 1.

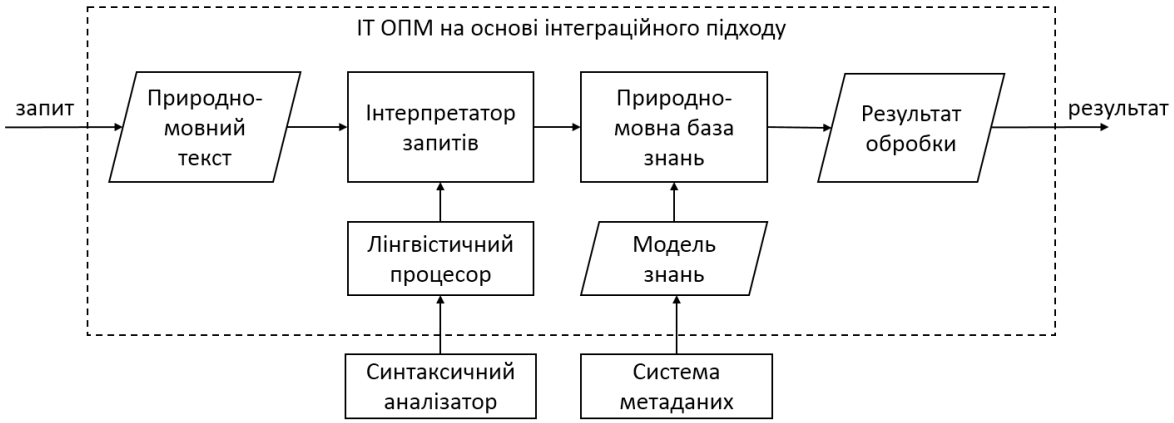

Рисунок 1 – Структурна схема інформаційної технології обробки природномовних текстів на основі інтеграційного підходу

Інтерпретатор запитів обробляє вхідні запити у вигляді природномовного тексту, виділяє їх структуру знань та перетворює їх на відповідні елементарні запити до ПМБЗ. В загальному випадку інтерпретатор підтримує базові запити (додавання, редагування, видалення та пошук у ПМБЗ), але за необхідності ця функціональність може бути розширена специфічними запитами, орієнтованими на певну прикладу задачу ОПМ. Ця підсистема є єдиною точкою входу в ІТ ОПМ; тобто, будь-які запити надходять виключно через інтерпретатор.

Лінгвістичний процесор виділяє структуру знань природномовного тексту з використанням методу обробки природної мови, наведеному у підрозділі 2.5. У своїй роботі лінгвістичний процесор використовує сторонній засіб, який не входить у саму ІТ ОПМ – синтаксичний аналізатор, орієнтований на деяку мову, який надає необхідні для роботи лінгвістичного процесора дані.

Природномовна база знань зберігає та надає доступ до природномовних знань, які формують базу знань ІТ ОПМ. Наповнення ПМБЗ організовано у відповідності до моделі, описаній у підрозділі 2.4. У роботі з ПМБЗ за необхідності може використовуватися зовнішня система метаданих, яка виконує додаткову обробку результатів запитів до ПМБЗ – наприклад, відбирає з кожного набору знайдених результатів лише ті, які є найбільш релевантними.

Залежно від напряму обробки ПМ текстів, ІТ ОПМ може працювати у режимах аналізу та синтезу тексту. Відповідно, аналіз – це отримання з вхідного ПМ тексту структури знань і її подальша обробка, а синтез – формування з існуючої структури знань ПМ тексту.

На рис. 2 представлена UML діаграма послідовності обробки даних ІТ ОПМ в режимі аналізу ПМ тексту.

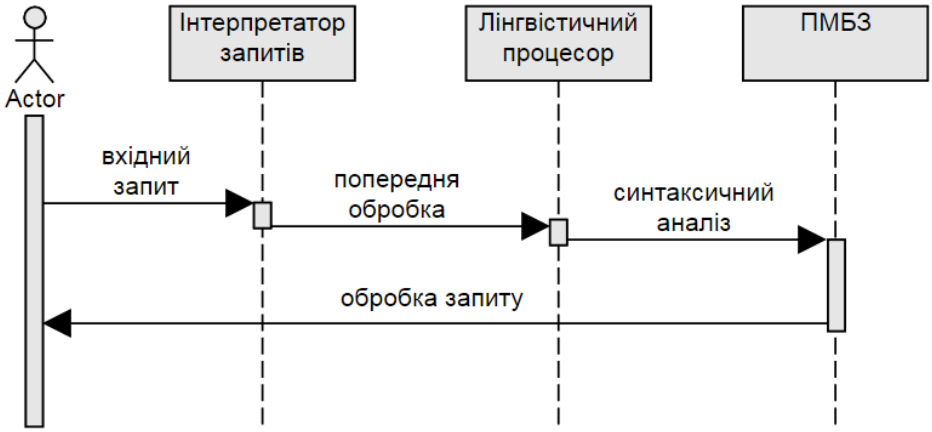

Рисунок 2 – Діаграма етапів роботи ІТ ОПМ в режимі аналізу природномовного тексту

При аналізі ПМ тексту виконуються наступні етапи обробки.

Етап 1. Попередня обробка.

Крок 1. Ідентифікація та виділення текстової інформації.

Крок 2. Видалення нетекстової інформації: формул, ілюстрацій тощо.

Крок 3. Розбивка вхідного тексту на менші фрагменти, як-то речення.

Етап 2. Синтаксичний аналіз. Виконується послідовно для кожного фрагменту.

Крок 1. Виділення слів та їх граматичних форм.

Крок 2. Виділення структури знань тексту.

Крок 3. Формування маркерів, які прив’язують структуру знань до синтаксичної структури ПМ тексту.

Етап 3. Обробка запиту.

Крок 1. Збереження отриманої структури знань у ПМБЗ.

Крок 2. Збереження фрагменту ПМ тексту у ПМБЗ.

Крок 3. Повернення користувачу посилання на фрагмент знань у ПБМЗ.

На рис. 3 представлена UML діаграма послідовності обробки даних ІТ ОПМ в режимі синтезу ПМ тексту.

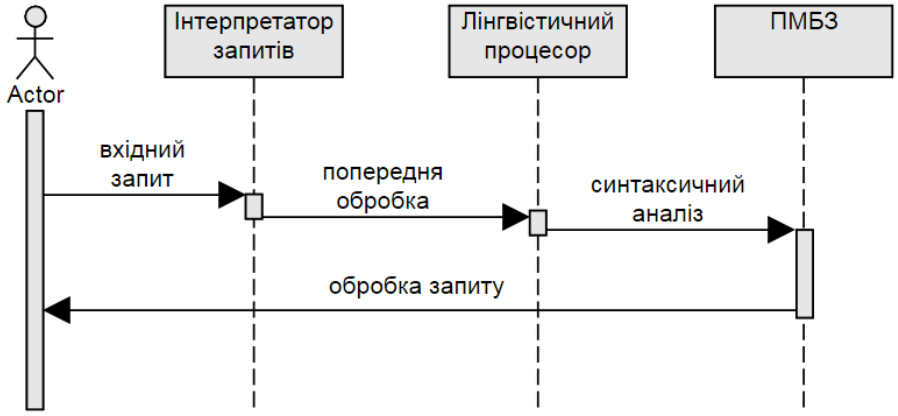

Рисунок 3 – Діаграма етапів роботи ІТ ОПМ в режимі аналізу природномовного тексту

При аналізі ПМ тексту виконуються наступні етапи обробки.

Етап 1. Попередня обробка.

Крок 1. Ідентифікація та виділення текстової інформації.

Крок 2. Видалення нетекстової інформації: формул, ілюстрацій тощо.

Крок 3. Розбивка вхідного тексту на менші фрагменти, як-то речення.

Етап 2. Синтаксичний аналіз. Виконується послідовно для кожного фрагменту.

Крок 1. Виділення слів та їх граматичних форм.

Крок 2. Виділення структури знань тексту з використанням метода, представленого у пунктах 2.5.1 – 2.5.3.

Крок 3. Формування маркерів, які прив’язують структуру знань до синтаксичної структури ПМ тексту.

Етап 3. Обробка запиту.

Крок 1. Збереження отриманої структури знань у ПМБЗ.

Крок 2. Збереження фрагменту ПМ тексту у ПМБЗ.

Крок 3. Повернення користувачу посилання на фрагмент знань у ПБМЗ.

Основними операціями над знаннями в ІТ ОПМ це пошук та записування знань, які є фактично базовими операціями записування/зчитування, адаптованими під формат знань ПМБЗ у складі ІТ.

Процедура обробки даних в ІТ ОПМ в режимі записування знань представлена на рис. 4.

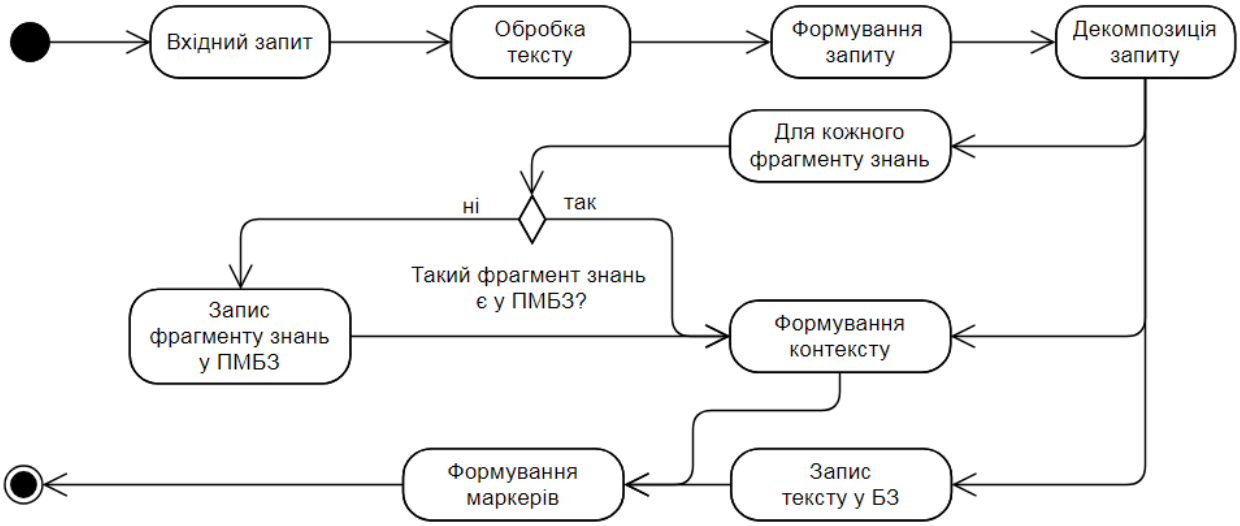

Рисунок 4 – Послідовність кроків записування знань в ІТ ОПМ

Процедура записування знань складається з наступних кроків.

1. Попередня обробка тексту. Вхідний запит у вигляді ПМ тексту перетворюється у фрагмент ПМ знань. Результатом виконання цього кроку є фрагмент знань, який відтворює структуру знань вхідного тексту.

2. Декомпозиція тексту. Структура знань, сформована на попередньому кроці, розбивається на окремі КЗ та відношення, при цьому окремо зберігаються зв’язки між ними та вхідний ПМ текст.

3. Наповнення ПМБЗ.

3.1. Додавання фрагментів знань у ПМБЗ. Для кожного отриманого фрагменту знань виконується перевірка наявності такого фрагменту у ПМБЗ. Якщо фрагмент не знайдено, він записується як новий елемент.

3.2. Додавання тексту у ПМБЗ. Для кожного фрагменту знань у ПМБЗ записується відповідна частина вхідного ПМ запиту.

4. Формування контексту. Після додавання усіх фрагментів вхідного запиту у ПМБЗ формуються зв’язки між ними відповідно до зв’язків, знайдених у вхідному тексті.

5. Формування маркерів. Після додавання усіх фрагменті знань до ПМБЗ також додаються зв’язки між знаннями та текстом на рівні лексем, квантів знань, фрагментів тексту та всього запиту відповідно. На цьому кроці за необхідності також додаються метадані, як-то джерело тексту, час та дата коли його було додано тощо.

Процедура обробки даних в ІТ ОПМ в режимі пошуку знань представлена на рис. 5.

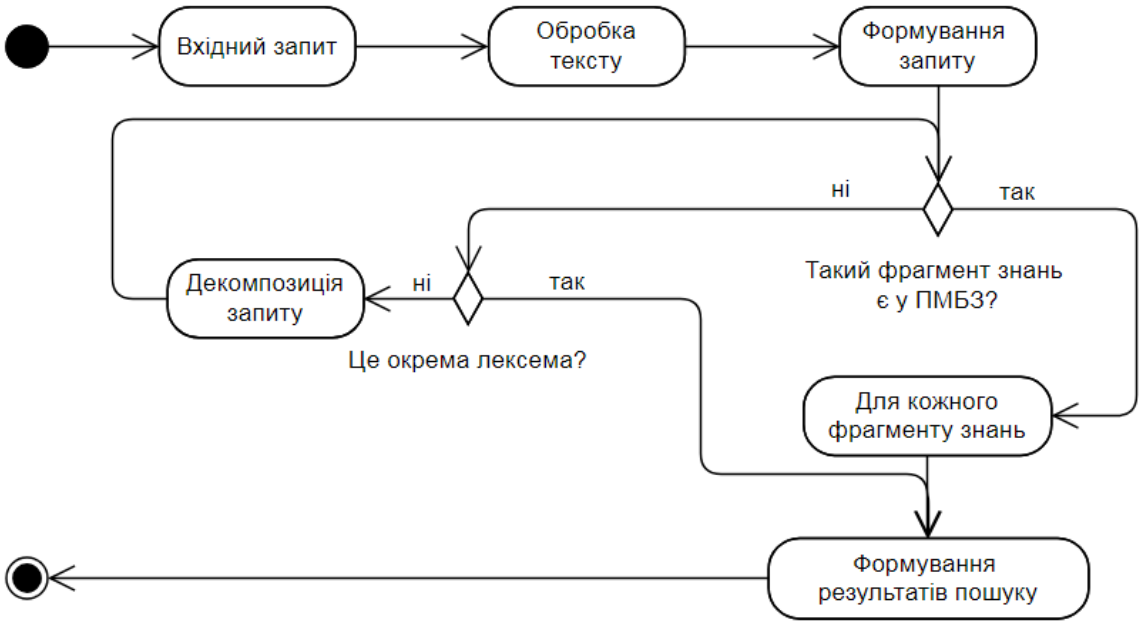

Рисунок 5 – Послідовність кроків пошуку знань в ІТ ОПМ

Процедура пошуку знань складається з наступних кроків.

1. Обробка тексту. Якщо пошуковий запит представлений у вигляді ПМ тексту, він перетворюється у фрагмент ПМ знань. В іншому випадку обробка починається безпосередньо з отриманого в запиті фрагменту знань.

2. Формування запиту. Виконується пошук даного фрагменту знань у ПМБЗ.

2.1. Якщо фрагмент знань знайдено повністю, виконується крок 3.

2.2. Якщо фрагмент знань не знайдено, виконується його декомпозиція на менші частини, до квантів знань та відношень включно, і для кожної частини виконується крок 2.

2.3. Якщо частину запиту зведено до окремої лексеми, виконується пошук за цією лексемою як пошук за ключовими словами, без врахування структури знань запиту.

3. Формування результату. Отримані результати сортуються у порядку, визначеному програмою, та повертаються до користувача. За необхідності результати доповнюються додатковими даними, як-то релевантність кожного результату, кількість входжень фрагменту знань у ПМБЗ тощо.

Також у роботі наведені процедури використання ІТ ОПМ у задачах природномовного пошуку та машинного перекладу.

Основними результатами природномовного пошуку з використанням ІТ ОПМ є ті документи, де лексеми пов’язані такими ж зв’язками, як і у пошуковому запиту. Це є перевагою при порівнянні з пошуком на основі ключових слів, оскільки пошук з використання ІТ ОПМ дозволяє відкинути документи, в яких лексеми запиту розташовані близько у тексті, але не пов’язані змістовно.

При виконанні машинного перекладу з використанням ІТ ОПМ запит розглядається як одне ціле, а переклад виконується на якомога більших його частинах, машинний переклад з використанням ІТ ОПМ дозволяє зберегти смислові зв’язки між частинами запиту, а в найкращому випадку навіть надати точний переклад запиту. Крім того, використання ПМБЗ в ролі джерела даних для мови-посередника дозволяє успішно перекласти синтаксичні варіації запиту, в яких смислові зв’язки залишаються незмінними.

У четвертому розділі визначено технічні вимоги до інформаційної системи, яка реалізує інформаційну технологію обробки природномовних знань на основі інтеграційного підходу (ІС ОПМ), зокрема визначено підсистеми та операції і розроблено схему бази даних для інформаційної системи. Проаналізовано обчислювальну складність пошуку природномовних знань у інформаційній системі та проведено її порівняння з аналогами. Виконано експериментальну перевірку використання інформаційної системи для підвищення релевантності результатів природномовного пошуку та виконано аналіз отриманих даних.

З огляду на особливості ІТ ОПМ для реалізації ІС ОПМ обрано веб-архітектуру, де основні компоненти інформаційної системи (а саме інтерпретатор запитів, ПМБЗ та лінгвістичний процесор) знаходяться на сервері додатків, що пов’язаний з сервером БД, користувачі підключаються до системи з використанням браузера або API, а системи синтаксичного аналізатора та метаданих взаємодіють з ІС ОПМ через Інтернет-підключення.

На основі розроблених моделі знань та методу обробки природномовних текстів виконано оцінку теоретичної складності пошуку в ІС ОПМ та її порівняння з аналогічними системами. Для цього визначено відносну складність пошуку як $K = \frac{C_2}{C_1}$, де C_2 – обчислювальна складність виконання задачі у ПМБЗ ІП, C_1 – відповідний показник її аналогу. Значення $K > 1$ показує відносну перевагу ПМБЗ ІП над аналогом, $K < 1$ – навпаки. Змінні, на основі яких виконується аналіз K – кількість елементів у БЗ та словнику і кількість слів у пошуковому запиті.

На основі даних про обсяг природномовних текстів та про кількість слів в окремих мовах визначено оцінку розміру словника ПМБЗ, що дорівнює 10^6 елементів. Для проведення аналізу виконано оцінку асимптотичного наближення для теоретично найбільшого значення розміру бази знань ($K = 10^{13}$).

Виконано аналіз обчислювальної складності для різних варіантів пошуку.

Пошук за рядком символів. Такий пошук відповідає операціям бінарного пошуку у масиві слів та масиві КЗ і, відповідно, має клас складності:

$$C = \log_2(N_w \cdot n), \tag{17}$$

де N_w – кількість слів у словнику;

n – кількість КЗ у БЗ.

Для найкращого аналогу, пошуку у БД за повнотекстовим індексом, за умови що середня довжина слова у різних мовах становить від 7 до 12 символів з медіаною близько 10 символів, складність пошуку становить:

$$C = \log_2(10 \cdot n).$$

Граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ становить:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(10 \cdot n)}{\ln(10^6 \cdot n)} = \frac{14}{19}. \tag{18}$$

Отже, для операції пошуку за символами гранична ефективність ПМБЗ становить ~74% від ефективності аналогів.

Пошук за складним запитом. Такий пошук виконується у декілька етапів, причому кількість етапів рівна кількості слів у запиті. Обчислювальна складність цієї операції складає:

$$C = \sum_{i=1}^{w} \log_2(N_i^Q \cdot N_i^R), \tag{19}$$

де N_i^Q – кількість КЗ у результатах пошуку на $(i-1)$ кроці,

N_i^R – кількість відношень кожного з цих КЗ,

w – кількість слів у запиті.

Найкращий аналог цієї операції – пошук у семантичних мережах, який реалізований як пошук у векторному просторі. Обчислювальна складність такого пошуку складає:

$$C = \log_2(N_Q \cdot N_L \cdot w), \tag{20}$$

де N_Q – загальна кількість записів у БЗ,

N_L – середня кількість словоформ у лексемі,

w – кількість слів у запиті.

Задача пошуку за складним запитом у ПМБЗ відповідає пошуку даного КЗ та його відношень у мережі КЗ. Складність такої операції, з урахуванням того, що в середньому КЗ має 7±2 відношення, складає:

$$C_1 = \sum_{n=0}^{6} 0{,}015^n \cdot \log_2(7 \cdot n) = 1{,}015 \cdot \log_2(7 \cdot n). \tag{21}$$

Для семантичних мереж складність цієї операції становить:

$$C_2 = \log_2(60 \cdot n). \tag{22}$$

Коефіцієнт відносної складності пошуку за символами визначається як:

$$K(n) = \frac{C_2}{C_1} = \frac{\log_2(60 \cdot n)}{1{,}015 \cdot \log_2(7 \cdot n)}. \quad (23)$$

Граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ дорівнює:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(60 \cdot n)}{1{,}015 \cdot \ln(7 \cdot n)} \approx 1{,}05162. \quad (24)$$

Отже, для операції пошуку за символами гранична ефективність ПМБЗ становить ~105% від ефективності аналогів.

Зазначимо, що при кількості записів на рівні мільйонів включно ($K < 10^6$) існуючі системи є достатньо потужними для виконання усіх перелічених задач, а отже – предметом інтересу є значення цього критерія для БЗ великого обсягу. На рис.6 наведено графік складності для обсягу БЗ на цій області визначення.

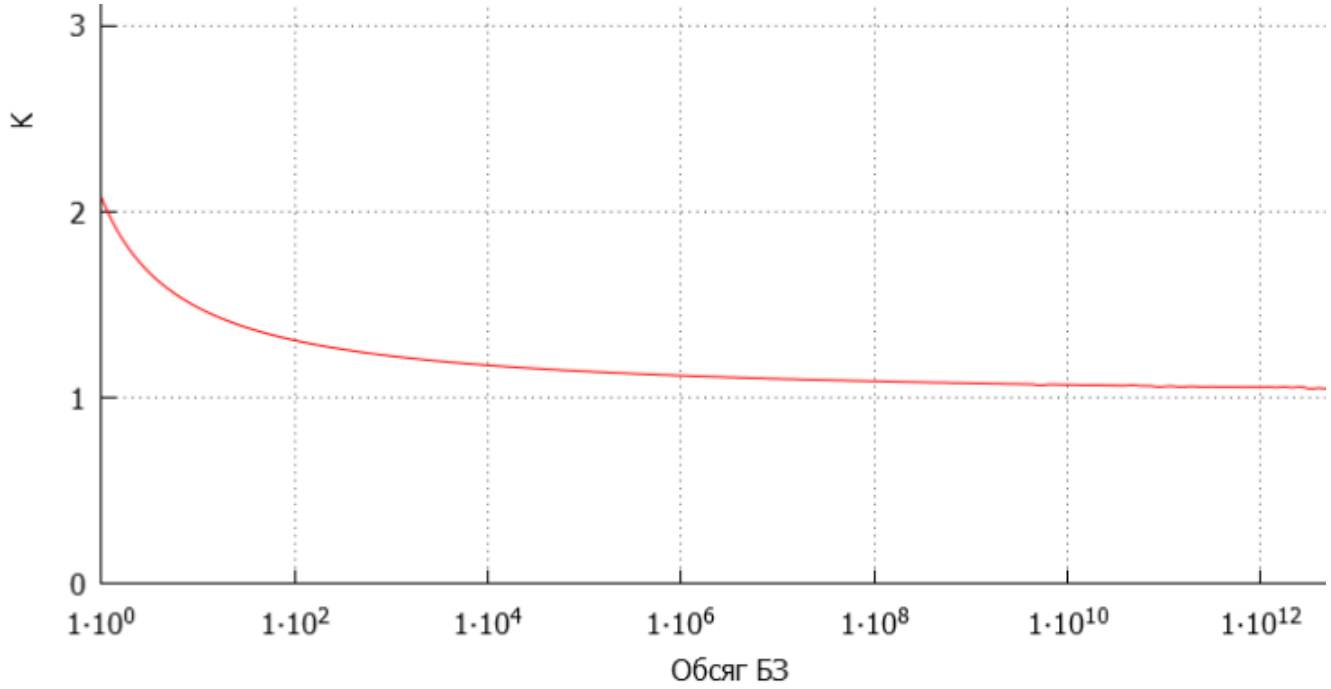

Рисунок 6 – Відносна складність пошуку, пошук за складним запитом,

Враховуючи, що кількість елементів, які відповідають вмісту пошукового запиту, значно менша за повне оточення для кожного з елементів K_S, а кількість таких з них, які одночасно відповідають ще й маркеру пошукового запиту (тобто складаються з заданих КЗ та відношень, поєднаних заданим чином у заданому порядку), ще звужує поле результатів пошуку, можемо стверджувати, що середня складність цієї операції буде значно менша за отриману складність.

Як видно з графіку, швидкість операції пошуку за складним запитом в ІС ОПМ перевищує таку в аналогів на значення від 5% до 12% залежно від обсягу БЗ.

В роботі було проаналізовано результати обробки пошукових запитів.

Як цільовий критерій було обрано релевантність пошуку – кількість результатів, релевантних пошуковому запиту, серед усіх знайдених результатів, у порівнянні з базовими результатами пошуку, тобто пошуком на основі ключових слів.

В ролі такого джерела запитів використовуємо набір різноманітних фрагментів тексту з природномовних джерел інформації, зокрема класичних творів української літератури, сайтів новин, юридичних документів, наукових публікацій. Вхідними даними для обробки було обрано перші 100 результатів (10 сторінок) з пошукової видачі природномовної пошукової системи *Google Search* для кожного з запитів.

Для кожного запиту було визначено та записано такі показники:

– текст запиту;
– кількість результатів пошуку;
– кількість результатів, які при обробці ІС ОПМ були автоматично визначені релевантними;
– помилкові результати – скільки результатів було помилково віднесено до релевантних або помилково пропущено;
– кількість результатів, які були правильно визначені релевантними.

Діапазон отриманих значень релевантності становить від 3% до 100%, тобто релевантність не обмежена жорсткими рамками, а ефективність додаткової обробки результатів пошуку ІС ОПМ залежить від запиту.

Для більшості запитів кількість помилкових результатів становить 0-1%, максимальне значення 6%. Таким чином, можемо стверджувати що вплив помилок обробки на її результати є відносно невеликим.

Середня релевантність до обробки результатів і після становить 72.5% та 72% відповідно. Це означає що помилкові запити незначно впливають на результати, і навіть при використанні примітивного лінгвістичного процесора похибка не є суттєвою.

Статистичні оцінки отриманих результатів: медіана дорівнює 86%, нижній квартиль 48.25%, верхній квартиль 97%. З цього можемо зробити такі висновки:

– додаткова обробки пошукових запитів з використанням ІС ОПМ є дуже ефективною для нижніх 25% результатів, де релевантність менша за 50%, і ідентифікація нерелевантних результатів значно зменшує обсяг даних. Це є важливим як для тих задач, які передбачають подальшу обробку цих даних, так і для пошуку за запитами які мають низьку початкову релевантність;
– обробка пошукових запитів з використанням ІС ОПМ не є ефективною для верхніх 25% результатів, релевантність яких сягає понад 97%. Цей показник достатньо хороший для більшості прикладних задач, і для його покращення краще застосовувати більш точні інструменти;
– обробки пошукових запитів з використанням ІС ОПМ дозволяє в середньому покращити ефективність на (100% - 86%) = 14%. Це достатньо хороший показник, який дозволяє стверджувати про ефективність роботи ІС ОПМ та доцільність її використання;
– ІС ОПМ може бути використана для оцінки релевантності результатів взагалі, зокрема при перевірці результатів перекладу.

ЗАГАЛЬНІ ВИСНОВКИ

Дисертаційна робота становить собою закінчене наукове дослідження, що вирішує актуальну науково-технічну задачу розробки інформаційної технології

обробки природномовних текстів на основі інтеграційного підходу. В рамках роботи поставлені та вирішені такі завдання:

1. На основі аналізу підходів до розробки природномовних баз знань обґрунтовано потребу у розробленні універсальної моделі представлення знань природномовного тексту для використання в технологіях обробки природної мови, що поєднує переваги існуючих підходів, та процедури використання такої моделі в інформаційних технологіях обробки природномовних текстів.

2. На основі інтеграційного підходу до моделювання мовленнєвої діяльності людини розроблено модель представлення знань для використання в технологіях обробки природної мови. Ця модель відрізняється від аналогів тим, що дозволяє представити фрагмент знань довільного природномовного тексту у вигляді універсальної структури. Перевагою розробленої моделі представлення знань є її незалежність від синтаксичної структури тексту та семантичного контексту фрагменту знань.

3. Розроблено процедури записування та пошуку знань з використанням розробленої моделі представлення знань в технологіях обробки природної мови, які дозволяють встановити зв'язки на структурному рівні між синтаксичною структурою тексту та довільною структурою метаданих.

4. Розроблено інформаційну технологію обробки природномовних текстів на основі інтеграційного підходу, для якої теоретично показано, що складність пошуку не перевищує так для аналогів, і в середньому є на 5-12% меншою для складних пошукових запитів.

5. Експериментально показано, що використання природномовної бази знань на основі інтеграційного підходу для природномовного пошуку дозволяє покращити якість роботи систем природномовного пошуку, а саме підвищити середню релевантність результатів на 14%. Впровадження результатів роботи у виробництві призвело до зменшення витрат часу працівників на контроль за складом продукції на 25% та збільшення конверсії природномовного пошуку на 8% відповідно.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

У виданнях іноземних держав:

1. Kyslenko Y, Sergeiev D. Cognitive architecture of speech activity and modelling thereof. Biologically Inspired Cognitive Architectures. 2015;12. *(Scopus; Elsevier, Нідерланди)* - *автору належить: модель бази знань, функціональна модель інтерфейсу.*

У наукових фахових виданнях України:

2. Кисленко ЮІ, Сергеєв ДС. Порівняння способів збереження слів в ІТ. Адаптивні системи автоматичного управління. 2016;(28(1)):33–41. *(OpenAIRE)* - *автору належить: формалізація проблеми, формалізація використаного підходу, збір та опрацювання даних.*

3. Кисленко ЮІ, Сергеєв ДС. Структурний підхід до пошуку природно-мовної

інформації. Радіоелектроніка та інформатика. 2015;(3):45–9. *(Index Copernicus)* - *автору належить: методика аналізу та опрацювання даних.*

4. Сергеєв ДС. A model of relation object for the natural language knowledge base [Модель об'єкту відношення для природно-мовної бази знань]. Адаптивні системи автоматичного управління. 2017;(30(1)):106–13. *(OpenAIRE)*

5. Сергеєв ДС, Хіміч АВ. Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях. Адаптивні системи автоматичного управління. 2016;(29(2)):140–6. *(OpenAIRE) – автору належить: формалізація та обгрунтування використаної моделі.*

6. Сергеєв ДС. Комп'ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань. Штучний інтелект. 2016;(4):42–8.

Матеріали конференцій:

7. Сергеєв ДС. Особливості моделювання бази природно-мовних знань. В: Електроніка та інформаційні технології ЕлІТ-2015. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2015. с. 17–20.

8. Сергеєв ДС. Методика оцінки якості роботи природно-мовних пошукових систем. In: SAIT-2016 [Internet]. Київ: НТУУ «КПІ», ІПСА; 2016. p. 413–6. Режим доступу: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf

9. Сергеєв ДС. Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції. В: ЕлІТ-2016. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2016. с. 25–8.

10. Сергеєв ДС. Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань. В: SAIT-2017. Київ: НТУУ «КПІ», ІПСА; 2017. с. 321–3.

11. Сергеєв ДС. Природно-мовна база знань як основа моделювання окремих аспектів мовленнєвої діяльності людини. В: Системи та засоби штучного інтелекту AIIS'2017. Київ: КНУ ім. Т.Шевченка, ф-т комп. наук та кібернетики; 2017. с. 171–8.

12. Сергеєв ДС. Integration model for knowledge representation for semantic WEB [Інтеграційна модель представлення знань для семантичного WEB]. В: ICSFTI2018. Київ: НТУУ «КПІ ім.Ігоря Сікорського», ф-т інформатики та обчислювальної техніки; 2018. с. 284–7.

АНОТАЦІЯ

Сергеєв Д.С. Інформаційна технологія обробки природномовних текстів на основі інтеграційного підходу. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» МОН України, м. Київ, 2019.

Дисертація присвячена вирішенню актуальної науково-технічної задачі розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

На основі аналізу актуальних проблем у галузі обробки природної мови показано, що прикладні технології обробки природної мови виконують поставлені задачі, але є можливим їх удосконалення для вирішення комплексних задач, зокрема машинного перекладу та природномовного пошуку. З цією метою створено формальну модель представлення знань у природномовній базі знань та моделі її основних елементів, якими є квант знань, або найменший елемент знань, та відношення, яке описує зв'язок між квантами знань. Розроблено метод обробки природномовних текстів на основі запропонованої моделі.

На основі створених моделей та методу розроблено процедури записування та пошуку природномовних знань для технологій обробки природної мови, які дозволяють встановити зв’язки на структурному рівні між синтаксичною структурою тексту та довільною структурою метаданих. Теоретично показано, що складність природномовного пошуку з використанням розроблених процедур не перевищує таку для аналогів, і в середньому є меншою ніж в аналогів для складних пошукових запитів.

В рамках роботи розроблено інформаційну технологію обробки природномовних текстів на основі інтеграційного підходу та експериментально показано, що її використання дозволяє підвищити середню релевантність природномовного пошуку на 14%.

Ключові слова: інформаційна технологія, природна мова, обробка природної мови, інтеграційний підхід, база знань, квант знань, пошук, машинний переклад.

АННОТАЦИЯ

Сергеев Д.С. Информационная технология обработки естественно-языковых текстов на основе интеграционного подхода. – На правах рукописи.

Диссертация на соискание научной степени кандидата технических наук по специальности 05.13.06 – информационные технологии. Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» МОН Украины, г. Киев, 2019.

Диссертация посвящена решению актуальной научно-технической задачи разработки информационной технологии обработки естественно-языковых текстов на основе интеграционного подхода.

На основе анализа актуальных проблем в области обработки естественного языка показано, что прикладные технологии обработки естественного языка выполняют поставленные задачи, но возможно их усовершенствования для решения комплексных задач, в частности машинного перевода и естественно-языкового поиска. С этой целью создана формальная модель представления знаний в естественно-языковой базе знаний и модели ее основных элементов, а именно кванта знаний, или наименьшего элемента знаний, и отношения, которое описывает связи между квантами знаний. Разработан метод обработки естественно-языковых текстов на основе предложенной модели.

На основе созданных моделей и метода разработаны процедуры записи и поиска естественно-языковых знаний для технологий обработки естественного языка, которые позволяют установить связи на структурном уровне между синтаксической

структурой текста и произвольной структурой метаданных. Теоретически показано, что сложность естественно-языкового поиска с использованием разработанных процедур не превышает таковую для аналогов, и в среднем является меньшей чем у аналогов для сложных поисковых запросов.

В рамках работы разработана информационная технология обработки естественно-языковых текстов на основе интеграционного подхода и экспериментально показано, что ее использование позволяет повысить среднюю релевантность естественно-языкового поиска на 14%.

Ключевые слова: информационная технология, естественный язык, обработка естественного языка, интеграционный подход, база знаний, квант знаний, поиск, машинный перевод.

ABSTRACT

Serheiev D.S. *Natural language texts processing technology based on the integrational approach.* – Manuscript.

Thesis for a Candidate Degree in Engineering, specialty 05.13.06 – information technologies. – National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute" of Ministry of Education and Science of Ukraine, Kyiv, 2019

In the recent years, the research in the field of natural language processing (NLP) has achieved significant practical results, including natural-language voice user interface for mobile devices, significant progress in machine translation technologies, handwriting and voice recognition, etc. At the same time, the task of improving the performance of these systems remains relevant. This study focuses on developing information technology for processing natural language texts based on the integrational approach, aimed to increase efficiency of natural language processing technologies. The subject of the research is models, methods, algorithms and information technologies for natural language processing.

Based on the analysis of actual problems in the field of natural language processing, it is shown that applied technologies of natural language processing are successful in fulfilling the intended specific tasks, but it is determined that there is a room for improvement in the area of solving complex problems, in particular, machine translation and natural language search. The role of knowledge bases in information technologies of natural language processing is determined as a necessary component for the interaction of different systems.

Existing approaches to developing natural-language knowledge bases are characterized and analyzed. A conclusion is made that existing technologies behind natural-language knowledge bases separately allow to achieve high levels of completeness, consistency and flexibility for practical purposes, but no technology combines high scores on all of the aforementioned qualities.

A formal model of knowledge representation for a natural-language knowledge base is created, including models of its main elements, namely the quantum of knowledge, or the smallest element of knowledge, and the relation objects that describe connections between quanta of knowledge. A method for processing natural language texts based on this model of knowledge is developed, including procedures for using information technology in applied natural language processing problems.

On the basis of the created models and the method, the procedures for writing and searching natural language skills for natural language processing technologies are developed, which allow to establish links at the structural level between the syntactic structure of the text and the arbitrary structure of the metadata. It is theoretically shown that the complexity of the natural-language search using the developed procedures does not exceed the complexity of the analogues, and on average is less than that of the analogues for complex search queries. Examples of use of the developed information technology for processing natural language texts in practical problems, namely natural language search and machine translation are provided. Writing and searching methods are created based on the knowledge representation model, allowing to establish links at the structural level between syntactic structure of the text and arbitrary structure of the metadata in natural language processing technologies.

Information technology for processing natural language texts based on the integrational approach is developed, for which it is theoretically proven that the search complexity does not exceed that of the existing alternatives, and is on average 5-12% lower for complex search queries. Subsystems and operations of such system are defined, and database scheme is developed. Computational complexity of natural language knowledge search in the information system is analyzed and compared with the existing alternatives.

Experimental testing of the information system is conducted and the acquired data are analyzed, demonstrating increased relevance of search results of natural language search. Within the framework of the work, information technology for the processing of natural language texts has been developed on the basis of the integrational approach. Based on experimental data acquired from measuring relevance of natural language search results, it has been shown that the developed information technology can increase relevance of search results. Specifically, relevance was increased by 14% on average for the whole set of experimental queries and search results, with no significant increase in relevance detected for the top quartile of results sorted by original relevance, and major increase detected for the lower quartile of original results.

The information technology for the processing of natural language texts can be used to improve performance of various natural language processing technologies, in particular natural language search systems, machine translation systems and natural language user interfaces.

Keywords: information technology, natural language, natural language processing, integrational approach, knowledge base, quantum of knowledge, search, machine translation.